Ensino & Informação
Prof.: Altamir A. R. Araldi
Revista®

Apresentação da Análise em Componentes Principais (ACP)

Programa

I. Objetivos da Análise em Componentes Principais

II. Comparação dos indivíduos e relação entre as variáveis na T. de D. na ACP

III Transformação da Matriz de Dados

IV A nuvem de pontos - indivíduos em $\mathbb{R}^3$ e em $\mathbb{R}^k$

V. Comparação dos indivíduos e forma da nuvem de pontos N^I

VI Busca de um novo referencial de representação da nuvem de pontos N^I

VII Cálculo dos eixos de inércia projetada máxima

VIII. As componentes principais

IX. Os planos principais

X. Qualidade Global de uma componente principal

XI. Contribuição do i - ésimo indivíduo à inércia projetada no comprimento do eixo α

XII. Qualidade de representação do i-ésimo indivíduo sobre o eixo α

XIII A nuvem de pontos - variáveis

Eduardo CRIVISQUI
Laboratoire de Méthodologie
du Traitement des Données
UNIVERSITÉ LIBRE DE BRUXELLES

I. Objetivos da Análise de Componentes Principais (ACP)

I.1. A Tabela de Dados

A ACP é utilizada para analisar a informação contida numa T. de D. de tipo indivíduos e variáveis quantitativas.

Tabla de Datos

(Individuos x variables cuantitativas)

	V_1	V_2	... V_k ...	V_K
1				
i			x_{ik}	
j			x_{jk}	
n				

Figura 1

I.2. Objetivos da ACP

→ O que significa analisar a informação...? :

- Avaliar a semelhança entre os indivíduos através dos atributos considerados:

  Existem grupos de indivíduos semelhantes...?

  Pode-se estabelecer uma topologia de indivíduos...?

- Avaliar a relação existente entre as características consideradas:

  Existem grupos de variáveis correlacionadas entre si...?

  Pode-se por em evidência uma topologia de variáveis...?

Nota : Para esta apresentação da Análise em Componentes Principais, baseámo-nos nos seguintes documentos :

- DROESBEKE, J-J. y FINE, J. *"Análisis de Componentes Principales"*, Programa PRESTA, 1995
- ESCOFIER, B. y PAGES, J. *"Análisis factoriales simples y múltiples - Objetivos, métodos e interpretación"* U.P.V., Bilbao, 1992.

II. Comparação dos indivíduos e relação entre as variáveis da T. de D. na ACP

II.1. Semelhança entre os indivíduos da T. de D.

Na ACP, dois indivíduos **i** e **j** são considerados mais semelhantes quanto maior for o número de valores similares apresentados no conjunto das variáveis.

A comparação dos indivíduos **i** e **j** é avaliada com a distância euclidiana clássica entre **i** e **j**:

$$d^2(i,j) = \sum_{k=1}^{K} m_k \left(x_{ik} - x_{jk} \right)^2$$

Sendo as variáveis consideradas com a mesma importância (isto é, o mesmo peso, $m_k = 1$),

$$d^2(i,j) = \sum_{k=1}^{K} \left(x_{ik} - x_{jk} \right)^2$$

II.2. Relação entre as variáveis da T. de D.

Na ACP, a relação entre as variáveis **k** e **p** é avaliada através do coeficiente de correlação (excepcionalmente, a covariância):

$$\bar{x}_k = \sum_{i=1}^{n} p_i x_{ik} \quad ; \quad r_{(k,p)} = \sum_{i=1}^{n} m_i \left(\frac{x_{ik} - \bar{x}_k}{s_{x_k}} \right) \left(\frac{x_{ip} - \bar{x}_p}{s_{x_p}} \right)$$

Como os indivíduos têm (normalmente) a mesma importância (o mesmo peso, $m_i = 1/n$):

$$r_{(k,p)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_{ik} - \bar{x}_k}{s_{x_k}} \right) \left(\frac{x_{ip} - \bar{x}_p}{s_{x_p}} \right)$$

III. Transformação da Matriz de Dados

Matriz de Datos

$$\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & x_{ij} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nk} \end{bmatrix}$$

$\Longrightarrow$

Matriz de Datos Centrada-Reducida

$$\begin{bmatrix} \frac{x_{11}-\bar{x}_1}{s_{x_1}} & \cdots & \frac{x_{1j}-\bar{x}_j}{s_{x_j}} & \cdots & \frac{x_{1k}-\bar{x}_k}{s_{x_k}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \frac{x_{ij}-\bar{x}_j}{s_{x_j}} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{s_{x_1}} & \cdots & \frac{x_{nj}-\bar{x}_j}{s_{x_j}} & \cdots & \frac{x_{nk}-\bar{x}_k}{s_{x_k}} \end{bmatrix}$$

Efeito da transformação dos dados:

- A avaliação da distância entre dois indivíduos e da correlação entre duas variáveis quaisquer da T. de D. não se muda pela operação de centragem da Matriz de Dados.

- A avaliação da relação entre duas variáveis quaisquer da T. de D. não se muda pela operação de redução da Matriz de Dados.

- A redução da Matriz de Dados faz com que a avaliação da semelhança entre dois indivíduos quaisquer da T. de D. seja independente das escalas de medida das variáveis.

IV. A nuvem de pontos – indivíduos em $\mathbb{R}^3$ e em $\mathbb{R}^k$

Seja uma nuvem de pontos de n indivíduos em R^3, dotada de uma métrica euclidiana.

A base (e_1,e_2,e_3) é um **base ortonormada**, centrada em G.

Ao eixo gerado por $e_1 = (1,0,0)$ corresponde a variável x_1 ... e assim por diante..

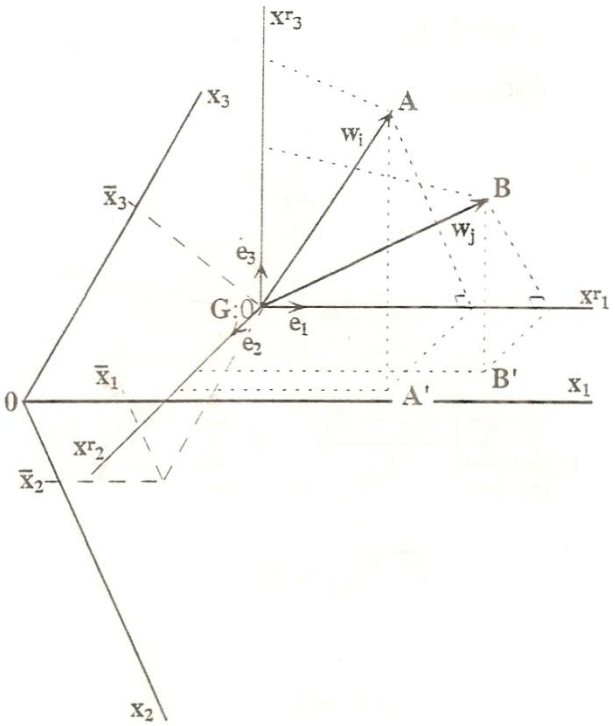


Figura 2

Um indivíduo **i** é representado da seguinte forma:

- O ponto **A** de coordenadas: $\left(\frac{x_{i1} - \bar{x}_1}{s_{x_1}}, \frac{x_{i2} - \bar{x}_2}{s_{x_2}}, \frac{x_{i3} - \bar{x}_3}{s_{x_3}} \right)$

- O extremo do vetor $\mathbf{W}_i$, combinação linear dos vetores da base ortonormada,

$$w_i = \left(\frac{x_{i1} - \bar{x}_1}{s_{x_1}} \right).e_1 + \left(\frac{x_{i2} - \bar{x}_2}{s_{x_2}} \right).e_2 + \left(\frac{x_{i3} - \bar{x}_3}{s_{x_3}} \right).e_3 = \begin{pmatrix} w_{i1} \\ w_{i2} \\ w_{i3} \end{pmatrix}$$

IV.1.Origem do espaço

O ponto **0** representa o "indivíduo médio".

No espaço original, o ponto **O** é o extremo do "vetor de médias" de todas as variáveis.

IV.2. Inércia total da nuvem de pontos – indivíduos

- Se representarmos o indivíduo **i** por um ponto no espaço R^3

dispersión de N_G^I = Inercia total de N_0^I

$$I_0^{N^I} = \frac{1}{n}\sum_{i=1}^{n} d^2(i,0)$$

Sendo : $d^2(i,0) = \sum_{k=1}^{3}\left(\frac{x_{ik}-\bar{x}_k}{s_{x_k}}\right)^2$

$$I_0^{N^I} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{3}\left(\frac{x_{ik}-\bar{x}_k}{s_{x_k}}\right)^2 = \sum_{k=1}^{3}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ik}-\bar{x}_k}{s_{x_k}}\right)^2$$

$$= \sum_{k=1}^{3}\frac{1}{s_{x_k}^2}\left(\frac{1}{n}\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2\right)$$

$$= \sum_{k=1}^{3}\frac{1}{s_{x_k}^2}\left(s_{x_k}^2\right) = k = 3$$

- Se representarmos o indivíduo i como extremo do vetor $\mathbf{w}_i$

$$I_0^{N^I} = \frac{1}{n}\sum_{i=1}^{n}\|w_i\|^2 \quad , \text{ siendo } \|w_i\|^2 = \sum_{k=1}^{3}\left(\frac{x_{ik}-\bar{x}_k}{s_{x_k}}\right)^2$$

$$I_0^{N^I} = Tr(V) = k = 3$$

- Em R^k :

$$I_0^{N^I} = Tr(V) = k$$

Como será definida a matriz **V**...?

Se chamarmos:

$\mathbf{I_k}$: matriz em $\mathbb{R}^k$

$$\underset{(k\,x\,k)}{I_k} = \begin{pmatrix} 1 & 0 & \vdots & & 0 \\ 0 & \ddots & \vdots & & \\ \cdots & \cdots & 1 & \cdots & \cdots \\ & & \vdots & \ddots & 0 \\ 0 & & \vdots & 0 & 1 \end{pmatrix}$$

D : matriz dos pesos em $\mathbb{R}^n$

$$\underset{(n\,x\,n)}{D} = \begin{pmatrix} \frac{1}{n} & 0 & \vdots & & 0 \\ 0 & \ddots & \vdots & & \\ \cdots & \cdots & \frac{1}{n} & \cdots & \cdots \\ & & \vdots & \ddots & 0 \\ 0 & & \vdots & 0 & \frac{1}{n} \end{pmatrix}$$

X : matriz de dados centrada-reduzida

$$\underset{(n\,x\,k)}{X} = \begin{pmatrix} & & & \vdots & & & \\ & & & \vdots & & & \\ \cdots & \cdots & \cdots & \frac{x_{ij} - \overline{x}_j}{s_{x_j}} & \cdots & \cdots & \cdots \\ & & & \vdots & & & \\ & & & \vdots & & & \end{pmatrix}$$

A matriz V é a matriz de correlações:

$$\underset{(k\,x\,k)}{V} = \underset{(k\,x\,n)}{X'}\ \underset{(n\,x\,n)}{D}\ \underset{(n\,x\,k)}{X} = \begin{pmatrix} 1 & r_{x_1x_2} & \vdots & & r_{x_1x_k} \\ r_{x_2x_1} & \ddots & \vdots & & \\ \cdots & \cdots & 1 & \cdots & \cdots \\ & & \vdots & \ddots & r_{x_jx_k} \\ r_{x_kx_1} & & \vdots & r_{x_kx_j} & 1 \end{pmatrix}$$

IV.3. Contribuição à inércia do indivíduo i

$$\text{CONTR.}\, I_0(i) = \frac{\frac{1}{n}\|w_i\|^2}{I_0} \times 100$$

IV.4 .Projeção ortogonal da nuvem de pontos sobre um eixo

Projetando ortogonalmente os pontos da nuvem N^I sobre o primeiro eixo, a inércia projetada é definida assim:

$$\hat{I}_0^1 = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{w}_i^1\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{i1}-\bar{x}_1}{s_{x_1}}\right)^2 = 1 = s_{x_1^r}^2$$

Se as variáveis são apenas centradas:

$$\hat{I}_0^1 = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{w}_i^1\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)^2$$

Igualmente para a projeção sobre o segundo eixo, o terceiro eixo,...

IV.5. Projeção ortogonal da nuvem de pontos sobre o plano definido pelos primeiros eixos

$$\hat{I}_0^{(1,2)} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{w}_i^{(1,2)}\right\|^2$$

$$\hat{I}_0^{(1,2)} = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_{i1}-\bar{x}_1}{s_{x_1}}\right)^2 + \left(\frac{x_{i2}-\bar{x}_2}{s_{x_2}}\right)^2\right] = 2 = s_{x_1^r}^2 + s_{x_2^r}^2$$

Se as variáveis são apenas centradas:

$$\hat{I}_0^{(1,2)} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{w}_i^{(1,2)}\right\|^2$$

$$\hat{I}_0^{(1,2)} = \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_{i1}-\bar{x}_1\right)^2 + \left(x_{i2}-\bar{x}_2\right)^2\right] = s_{x_1}^2 + s_{x_2}^2$$

Inércia projetada $\mapsto$ soma de variâncias $\mapsto$ "dispersão".

V. Comparação dos indivíduos e forma da nuvem de pontos N^I

☞ A **forma** da nuvem de pontos é dada pelo conjunto de distâncias entre os pontos - indivíduos.

Para fazer o balanço dessas distâncias deve-se estudar a forma da nuvem de pontos - indivíduos...identificar uma partição dos pontos, ou as direções de alongamento dessa nuvem de pontos.

☞ Se o número de variáveis é superior a 3, o estudo direto da forma da nuvem de pontos – indivíduos em R^k ($k>3$) é impossível.

☞ Deve-se buscar imagens planas que representem "o melhor possível" a disposição dos pontos no espaço $\mathbb{R}^k$.

VI. Busca de um novo referencial de representação da nuvem de pontos N^I

*** Problema :**

- Dispõe-se de uma nuvem de pontos em $\mathbb{R}^k$.

Busca-se uma série $\{u_s ; s = 1, 2, \ldots, S\}$ de direções privilegiadas de R^k, chamadas "direções principais de alongamento" da nuvem de pontos, tais que tomadas duas a duas definam os planos principais sobre os quais se projeta a nuvem de pontos.

Essas direções principais devem ser tais que:

☞ Cada direção principal da série seja ortogonal às direções definidas precedentemente,

$$\left\{ u_s ; s = 1, 2, \ldots, S \,\middle|\, u_2 \perp u_1 \; ; \; u_3 \perp u_2 \perp u_1 \; ; \; \ldots \; ; \; u_s \perp \ldots \perp u_3 \perp u_2 \perp u_1 \right\}$$

☞ Cada direção principal u_s maximize a inércia com respeito à origem da projeção da nuvem de pontos ao longo de $\mathbf{u}_s$.

☞ O plano definido por (u_1, u_2) maximize a inércia com respeito à origem da projeção da nuvem de pontos sobre esse plano.

☞ O subespaço definido por $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ maximize a inércia com respeito à origem da projeção da nuvem de pontos sobre esse subespaço....

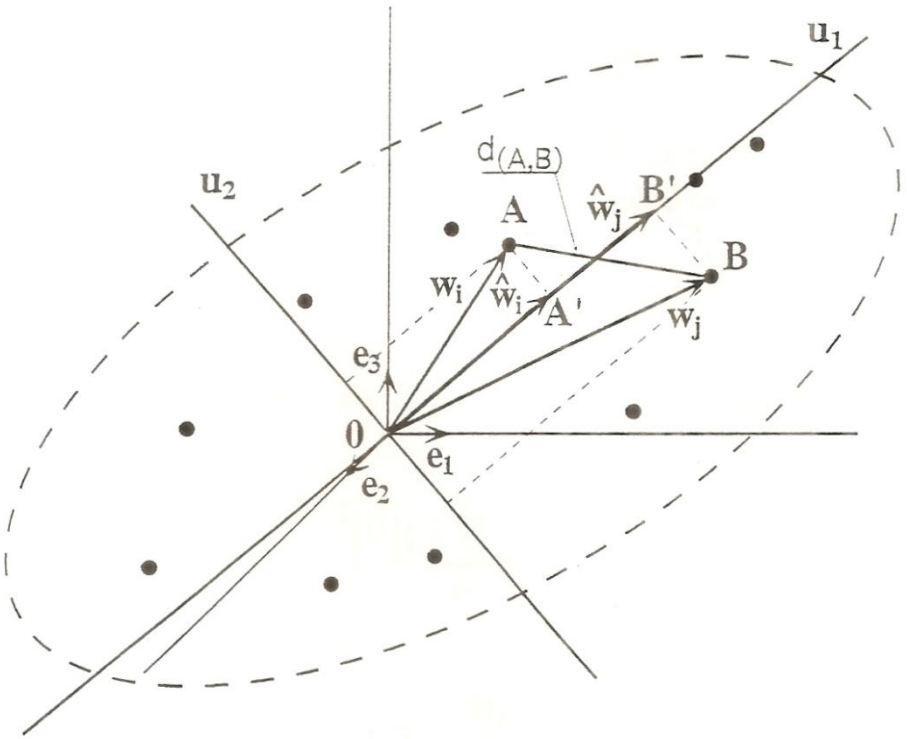


Figura 3

☞ É equivalente buscar:

- A direção $\mathbf{u}_1$ que maximiza :

$$\hat{I}_0^{(u_1)} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{w}_i^{(1)}\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}|0A'|^2$$

- A direção $\mathbf{u}_1$ que minimiza :

$$\hat{I}_0^{(u_2)} = \frac{1}{n}\sum_{i=1}^{n}\left\|w_i - \hat{w}_i^{(1)}\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}|AA'|^2$$

- Interpretação das direções principais:

■ Em razão da centralização da série $\{u_s; s = 1, 2, \ldots, S\}$, as direções principais podem ser interpretadas como as direções de alargamento máximo da nuvem de pontos – indivíduos.

As direções principais são eixos que representam "o melhor" possível a variabilidade (diversidade) dos indivíduos.

■ Pode-se demonstrar que, em razão da centralização da série $\{u_s; s = 1, 2, \ldots, S\}$, maximizar:

$$\hat{I}_0^{(u_1)} = \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{w}_i^{(1)} \right\|^2$$

É equivalente a maximizar:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \hat{w}_i^{(k)} - \hat{w}_j^{(k)} \right\|^2 \quad , \forall\ k = 1, 2, \ldots, S$$

A projeção da nuvem $\mathbf{N^I}$ ao longo de uma direção principal reduz as distâncias entre os pontos da nuvem.

Mas as direções principais são tais que as distâncias entre os pontos da nuvem $\mathbf{N^I}$ são "as mais próximas" possíveis das distâncias entre os pontos dessa nuvem no espaço original.

Mas, como definir a série $\{u_s; s = 1, 2, \ldots, S\}$ com essas propriedades...?

VII. Cálculo dos eixos para inércia projetada máxima

O problema se resume na busca de um vetor $w \in \mathbb{R}^k$ tal que :

- $w' w = 1$

- e que faça máxima a projeção w'X'DXw, isto é a projeção ao longo da diração w da nuvem de pontos N^I.

☞ A solução é obtida pela "diagonalização" da matriz de inércia...

a) Diagonalização da matriz V

Os **p auto-valores** (característicos), não nulos, de uma matriz simétrica $V_{(k\times k)}$, (**p** ≤ **k**), denotados por $\{\lambda_\alpha, \alpha = 1,\ldots, p\}$, são as soluções da "equação característica":

$$\left| \underset{(k\times k)}{V} - \lambda_\alpha \underset{(k\times k)}{I} \right| = 0$$

Os **p auto-valores** não nulos λ_α e os (**k-p**) auto-valores nulos da matriz $V_{(k\times k)}$ são tais que anulam o determinante da matriz $(V - \lambda_\alpha I)$.

Nessa expressão, $I_{(k\times k)}$ é a matriz identidade.

A equação característica pode ser assim formulada:

$$\left| \underset{(k\times k)}{V} - \underset{(k\times k)}{D_{\lambda\alpha}} \right| = 0 \qquad (1)$$

A cada auto-valor λ_α da matriz $V_{(k\times k)}$ corresponde um auto-vetor u_α.

Os auto-vetores associados à matriz $V_{(k\times k)}$ estão definidos pelas soluções do sistema das seguintes equações lineares:

$$\left(\underset{(k\times k)}{V} - \underset{(k\times k)}{D_{\lambda\alpha}} \right) . u_\alpha = 0 \qquad (2)$$

Esse sistema de equações tem solução (trivial $u_{\alpha j} = 0 \ \forall \ \alpha = 1,\ldots, p \ ; \ y \ \forall j = 1,\ldots, k$) já que o determinante de $\left(V - D_{\lambda\alpha}\right)$ é nulo.

O sistema de equações comporta **k-1** equações lineares independentes com **k** variáveis desconhecidas, de modo que qualquer solução $u_\alpha = \left(u_{\alpha 1}, \ldots, u_{\alpha k}\right)$ só pode ser estabelecida com um fator constante de proporcionalidade.

Impondo uma condição suplementar:

$$u_{\alpha 1}^2 + u_{\alpha 2}^2 + \cdots + u_{\alpha k}^2 = 1$$

O sistema é levado a um sistema de k equações lineares independentes com k variáveis desconhecidas. A solução é única.

A cada auto-valor λ_α não nulo corresponde então uma solução não trivial u_α.

Define-se assim uma nova base ortogonal constituída pelos p auto-vetores u_α, mutuamente ortogonais e de norma $\|u_\alpha\|^2 = 1$.

Ordenando por ordem decrescente os auto-valores,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

E os auto-vetores associados, estabelece-se a matriz $U_{(k,k)}$ de faixa **p**, composta dos vetores próprios de $V_{(k,k)}$ em coluna, soluções do sistema de equações (2).

O sistema de equações (2) pode então ser assim formulado:

$$\underset{(k \times k)}{V} \cdot u_\alpha = \lambda_\alpha . u_\alpha \quad ; \quad \text{o bien} \quad \underset{(k \times k)(k \times k)}{V \ U} = \underset{(k \times k)}{U} \ \underset{(k \times k)}{D_{\lambda\alpha}} \qquad (3)$$

Sendo $U_{(k,k)}$ de faixa **p**, não se trata de uma matriz singular, de modo que se pode definir sua inversa $U^{-1}_{(k,k)}$.

Pós multiplicando em (3), obtém-se que:

$$\underset{(k\times k)}{V}\ \underset{(k\times k)}{U}\ \underset{(k\times k)}{U^{-1}} = \underset{(k\times k)}{U}\ \underset{(k\times k)}{D_{\lambda\alpha}}\ \underset{(k\times k)}{U^{-1}}$$

$$\Rightarrow \quad \underset{(k\times k)}{V} = \underset{(k\times k)}{U}\ \underset{(k\times k)}{D_{\lambda\alpha}}\ \underset{(k\times k)}{U^{-1}} \qquad (4)$$

Pré multiplicando em (3), obtém-se que:

$$\underset{(k\times k)}{U^{-1}}\ \underset{(k\times k)}{U}\ \underset{(k\times k)}{D_{\lambda\alpha}} = \underset{(k\times k)}{U^{-1}}\ \underset{(k\times k)}{V}\ \underset{(k\times k)}{U}$$

$$\Rightarrow \quad \underset{(k\times k)}{D_{\lambda\alpha}} = \underset{(k\times k)}{U^{-1}}\ \underset{(k\times k)}{V}\ \underset{(k\times k)}{U} \qquad (5)$$

Considerando as expressões (4) e (5):

Existe uma matriz $U_{(k\times k)}$ que permite transformar a matriz de inércia $V_{(k\times k)}$ em uma matriz diagonal $D^{\lambda_\alpha}_{(k\times k)}$, que é uma matriz semelhante a $V_{(k\times k)}$.

Levando a equação característica à forma:

$$\left(\underset{(k\times k)}{V} - \lambda_\alpha \underset{(k\times k)}{I}\right) u_\alpha = 0 \quad \Rightarrow V.u_\alpha = \lambda_\alpha . u_\alpha$$

Mostra-se que:

- u_α é uma direção privilegiada da matriz **V**,
- Já que $V.u_\alpha$ corresponde a uma multiplicação simples do vetor u_α por um escalar λ_α,
- A direção não se modifica e sua norma é multiplicada por λ_α.

A base ortonormada $\{u_\alpha\ ,\ \alpha = 1,\ldots, p\}$ define as direções principais de alongamento da nuvem $\mathbf{N^I}$.

Dessa forma, pode-se demonstrar que as direções principais são as direções de inércia máxima...

b) Eixos da inércia máxima

Existem :

- **p** valores reais positivos tais que:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

São os **valores próprios** de la matriz $X' D X = V$.

- os auto-vetores associados $\{u_\alpha , \alpha = 1,\ldots, p\}$ constituem uma base ortonormada de $\mathbb{R}^p$.

Decompondo um vetor **W** nesta base:

$$w = \sum_p w_p u_p \quad , \text{ con } \sum_p w_p^2 = 1$$

$$X' D X w = X' D X \sum_p w_p u_p = \sum_p \lambda_p w_p u_p$$

A inércia projetada sobre w pode ser escrita:

$$w' X' D X w = \sum_p \lambda_p w_p^2 \leq \lambda_1 \sum_p w_p^2 \qquad (6)$$

com a exigência: $\sum_p w_p^2 = 1$ 'a expressão (6) é máxima para λ_1.

Esse máximo é obtido quando: $w_1 = (1, 0;\ldots, 0)$, $w_1 \in \mathbb{R}^p$,

isto é, quando: $w = \pm u_1$

A inércia da nuvem de pontos N^I projectada ao longo de um eixo **w** é máxima quando este eixo é colinear ao vetor próprio u_1 de **V** asociado ao maior valor próprio λ_1.

VIII. As componentes principais

As componentes principais são obtidas pela projeção ortogonal da nuvem de pontos $\mathbf{N}^{\mathbf{I}}$ *ao longo de cada direção principal.*

A coordenada do **i**-ésimo *indivíduo no novo sistema de eixos* (origem em **G**, base ortonormada $\{u_\alpha\ ,\ \alpha = 1,\ldots, p\}$) *está definida pelo produto escalar* $<\mathbf{w}_i, \mathbf{u}_\alpha>$.

$$F_\alpha(i) = \sum_{j=1}^{p} u_{\alpha j}\left(\frac{x_{ij} - \bar{x}_j}{s_{x_j}}\right)$$

Verifica-se que:

- *As componentes principais são variáveis centradas,*

$$\bar{x}_\alpha = \frac{1}{n}\sum_{i=1}^{n} F_\alpha(i) = 0$$

- *A variância da componente principal* α, *é o auto-valor* λ_α,

$$\mathrm{var}_\alpha = \frac{1}{n}\sum_{i=1}^{n} F_\alpha^2(i) = \lambda_\alpha$$

- *As componentes principais são variáveis não correlacionadas,*

$$\mathrm{COV}\left(F_p, F_k\right) = 0$$

IX. Os planos principais

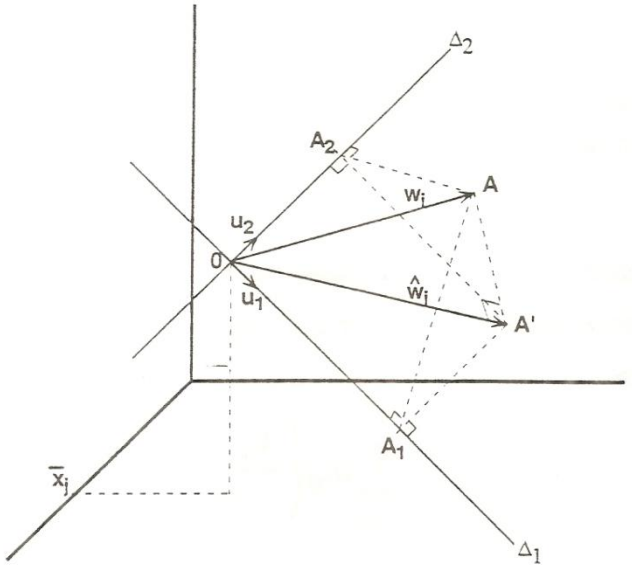


Figura 4

$\mathbf{A}_1$: *projeção ortogonal do ponto* A, sobre Δ_1

$\mathbf{A}_2$: *projeção ortogonal do ponto* A, sobre Δ_2

$\mathbf{A}'$: *projeção ortogonal do ponto* A, sobre o plano (Δ_1, Δ_2)

Pelo teorema de Pit goras:

$$d^2(O,A) = d^2(O,A') + d^2(A,A')$$
$$= d^2(O,A_1) + d^2(O,A_2) + d^2(A,A')$$

De modo que:

$$I_0^{N^I} = \frac{1}{n}\sum_{i=1}^{n} |O,A_i|^2$$

$$I_0^{N^I} = \frac{1}{n}\sum_{i=1}^{n} |O,A_{1i}|^2 + \frac{1}{n}\sum_{i=1}^{n} |O,A_{2i}|^2 + \frac{1}{n}\sum_{i=1}^{n} |A_i,A'_i|^2$$

$$I_0^{N^I} = \frac{1}{n}\sum_{i=1}^{n} F_1^2(i) + \frac{1}{n}\sum_{i=1}^{n} F_2^2(i) + \underline{\frac{1}{n}\sum_{i=1}^{n} F_3^2(i)} \quad \text{mínimo}$$

$$I_0^{N^I} = \lambda_1 + \lambda_2 + \lambda_3 = I_{(1,2)}^{N^I} + I_3^{N^I}$$

X. Qualidade global de uma componente principal

Generalizando...em $\mathbb{R}^p$

w_i: *vetor cuja extremidade representa o* i-ésimo *ponto–indivíduo.*

$F_\alpha(i)$: *longitude da projeção de* w_i *ao longo do eixo* α.

$$I_0^{N^I} = \frac{1}{n}\sum_i \|w_i\|^2 = \frac{1}{n}\sum_i F_1^2(i) + \ldots + \frac{1}{n}\sum_i F_p^2(i)$$

$$I_0^{N^I} = \quad k \quad = \quad \lambda_1 \quad + \ldots + \quad \lambda_p$$

Pode-se avaliar a qualidade global da α-èsima componente principal :

$$\tau_\alpha = \frac{\lambda_\alpha}{I_0^{N^I}} \times 100 = \frac{\lambda_\alpha}{k} \times 100$$

XI. *Contribuição do i-ésimo indivíduo à inércia projetada ao longo do eixo* α

Sendo:

$$I_\alpha^{N^I} = \frac{1}{n}\sum_i F_\alpha^2(i) = \lambda_\alpha$$

A contribuição do i-ésimo indivíduo à inércia projetada ao longo do eixo α :

$$\mathrm{CONTR}_\alpha(i) = \frac{\frac{1}{n}F_\alpha^2(i)}{\lambda_\alpha} \times 100$$

XII. Qualidade de representação do i-ésimo indivíduo sobre o eixo α

$$\cos^2_{\alpha}(w_i, u_i) = \frac{F_{\alpha}^2(i)}{\|w_i\|^2}$$

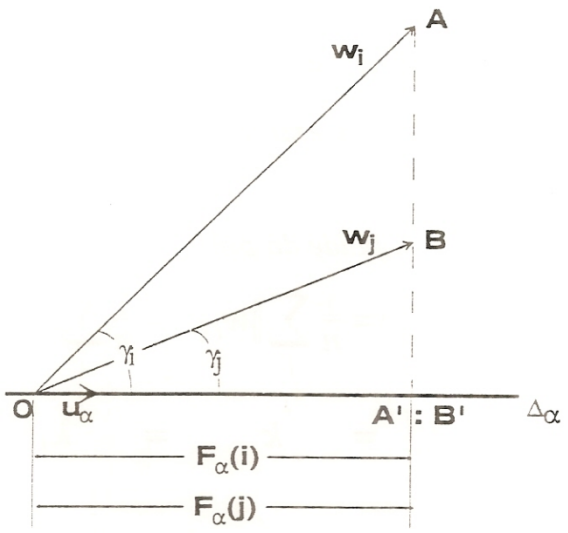


Figura 5

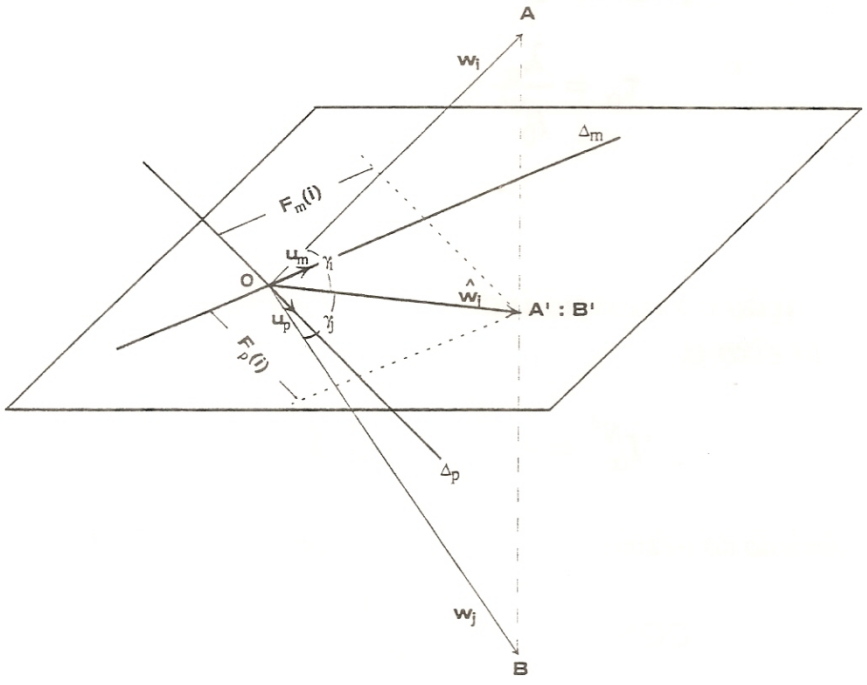


Figura 6

$$\cos^2_{(m,\,p)}(w_i, \hat{w}_i) = \frac{F_m^2(i) + F_p^2(i)}{\|w_i\|^2}$$

XIII. A nuvem de pontos – variáveis

- Uma variável é representada por um vetor em $\mathbb{R}^n$.

- A métrica em $\mathbb{R}^n$ *é dada pela matriz* $\mathbf{D}_{(n \times n)}$. *Isto é, todos os indivíduos têm o mesmo peso* ($p_i = 1/n$).

- *Produto escalar de duas variáveis:*

$$\langle x_j, x_k \rangle = \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik}$$

- *Interpretação geométrica dos principais índices estatísticos:*

Seja: $x_{ij}^r = \dfrac{x_{ij} - \bar{x}_j}{s_j}$, la variável centrada e reduzida corresp. a la var. $\mathbf{x}_j$.

Seja: $1 = (1,..,1)$, un vector de $\mathbb{R}^n$ con todos sus elementos iguais a 1.

* *Média* $\mathbf{D}^{\lambda_\alpha}_{(k \times k)(k \times k)}$

* *Covariância:*

$$\langle (x_j - \bar{x}_j), (x_k - \bar{x}_k) \rangle = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = COV(x_j, x_k)$$

* *Variância*

$$\| (x_j - \bar{x}_j) \|^2 = \langle (x_j - \bar{x}_j), (x_j - \bar{x}_j) \rangle = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 = V(x_j) = s_{x_j}^2$$

* *Desvio padrão* $\| (x_j - \bar{x}_j) \| = s_{x_j}$

* *Correlação*

$$\cos\left((x_j - \bar{x}_j), (x_k - \bar{x}_k) \right) = \frac{\langle (x_j - \bar{x}_j), (x_k - \bar{x}_k) \rangle}{\| (x_j - \bar{x}_j) \| \cdot \| (x_k - \bar{x}_k) \|} = \frac{COV(x_j, x_k)}{s_{x_j} \cdot s_{x_k}} = r_{(x_j, x_k)}$$

- *O conjunto de extremidades dos vetores que representam as variáveis constituem a nuvem de pontos* $\mathbf{N}^k$.

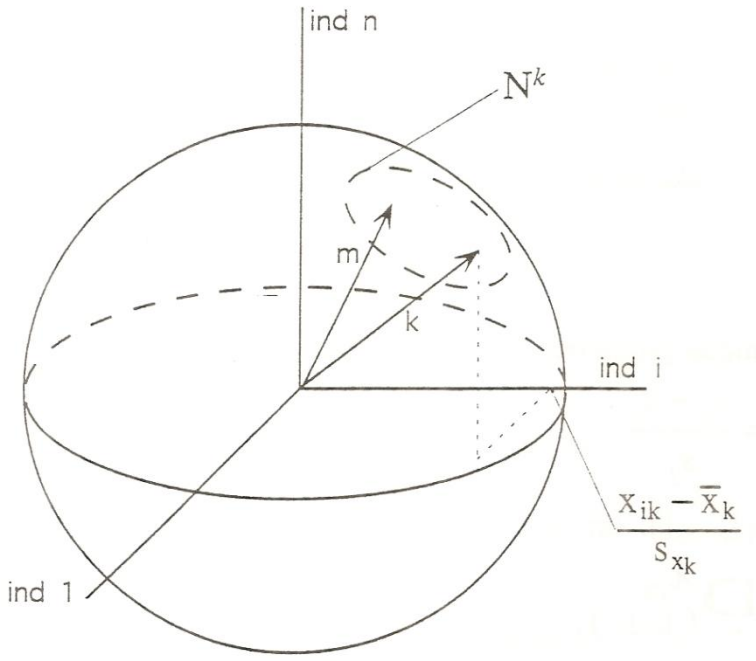


Figura 7

$$\|m\|^2 = \frac{1}{n}\sum_i \left(\frac{x_{im} - \bar{x}_m}{s_{x_m}}\right)^2 = 1 = \|k\|^2$$

- A nuvem de pontos N^k *está situada em uma hiperesfera de raio 1.*

$$- \cos(m,k) = < m,k > = \frac{1}{n}\sum_i \left(\frac{x_{im} - \bar{x}_m}{s_{x_m}}\right)\left(\frac{x_{ik} - \bar{x}_k}{s_{x_k}}\right)$$

$$= \mathrm{corr}(m,k) = r_{m,k}$$

Como a norma dos vetores que representam as variáveis igual a 1, a coordenada da projeção de uma variável sobre outra pode ser interpretada como o coeficiente de correlação entre as variáveis.

☞ Fazer o balanço dos coeficientes de correlação entre as variáveis equivale a estudar os ângulos entre os vetores que definem a nuvem $\mathbf{N}^k$.

O estudo direto é impossível em razão das dimensões de $\mathbb{R}^n$.

A ACP produz as variáveis que constituem um resumo das variáveis iniciais e que permitem a representação plana aproximada das variáveis e de seus ngulos respectivos.

XIII.1. *A ACP em* $\mathbb{R}^n$, *espaço das variáveis*

Disp e-se das seguintes matrizes:

X : *matriz de dados centrados - reduzidos*

D : *métrica de pesos em* $\mathbb{R}^n$

I : *métrica de* $\mathbb{R}^k$

Pode-se definir as direções principais s_α *tais que:*

$$XIX'Ds_\alpha = \lambda_\alpha s_\alpha \quad \text{, sendo } \|s_\alpha\| = 1$$

- A primeira componente principal c_1 *é a combinação linear das* **k** *variáveis de* **X** *que têm variância máxima.*

- A segunda componente principal c_2 *é a combinação linear das* **k** *variáveis de***X**, *ortogonal à primeira componente e que têm variância máxima.*

- E assim segue..... *As componentes principais* (c1,..., ck) *formam uma base ortogonal de* $\mathbb{R}^n$ de **k** *dimensões, definidas pelas* **k** *variáveis.*

$\{s_\alpha ; \alpha = 1,\ldots,k\}$: *base canônica do sub- epaço de* $\mathbb{R}^n$ *de dimensão* **k.**

Seja $v_\alpha = \dfrac{c_\alpha}{\sqrt{\lambda_\alpha}} \Rightarrow \{v_\alpha ; \alpha = 1,\ldots,k\}$: *base ortonormal do subespaço de* $\mathbb{R}^n$.

Se obtém:

$$x_j^r = \sum_{\alpha=1}^{k} r_{(x_j, c_\alpha)} v_\alpha ; j = 1,\ldots,k$$

As k vari veis centradas-reduzidas s o vetores cujas extremidades se colocam sobre a esfera de raio 1.

XIII.2. *Projeção de x_j^r sobre o primeiro plano fatorial, gerado por v_1, v_2*

Seja $\hat{x}_j^r$ a projeção de x_j^r sobre o primeiro plano fatorial definido por c_1 e c_2, obtém-se:

$$\hat{x}_j^r = \sum_{\alpha=1}^{2} r_{(x_j, c_\alpha)} v_\alpha ; j=1,\ldots,k$$

As coordenadas das vari veis centradas e reduzidas sobre o primeiro plano fatorial s o as correlaç es das vari veis com as componentes principais.

XIII.3. *Qualidade de representação de uma variável*

Traçando o círculo de raio 1 no primeiro plano fatorial, podemos avaliar visualmente a qualidade de representação de uma variável x_j^r mediante $\hat{x}_j^r$.

Já que $\|x_j^r\| = 1$, se a extremidade de $\hat{x}_j^r$ se colocar próxima do círculo de raio 1, também $\|\hat{x}_j^r\|$ será próxima de 1, isto é considera-se que x_j^r tem uma boa qualidade de representação.

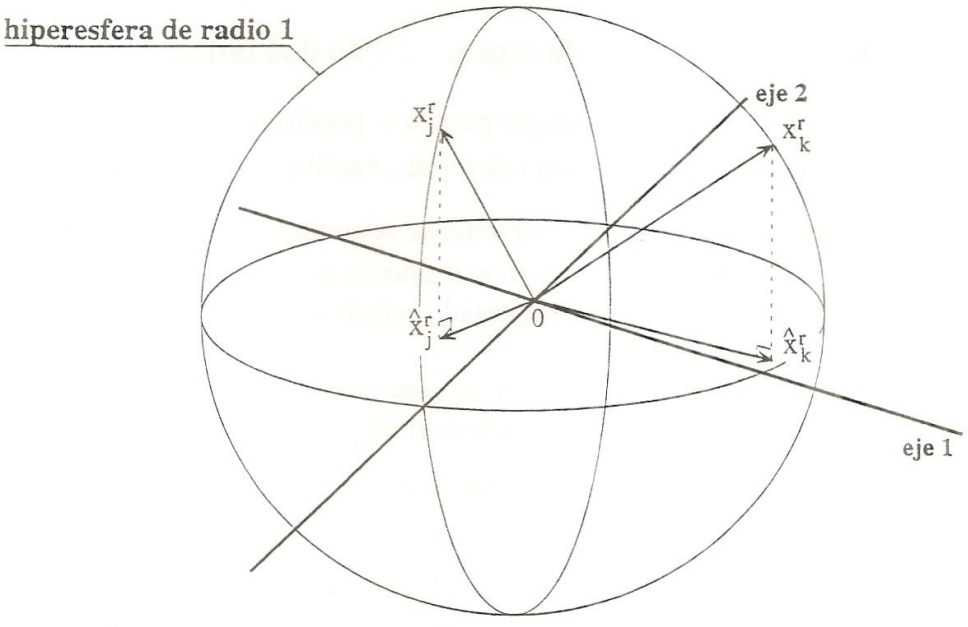


Figura 8

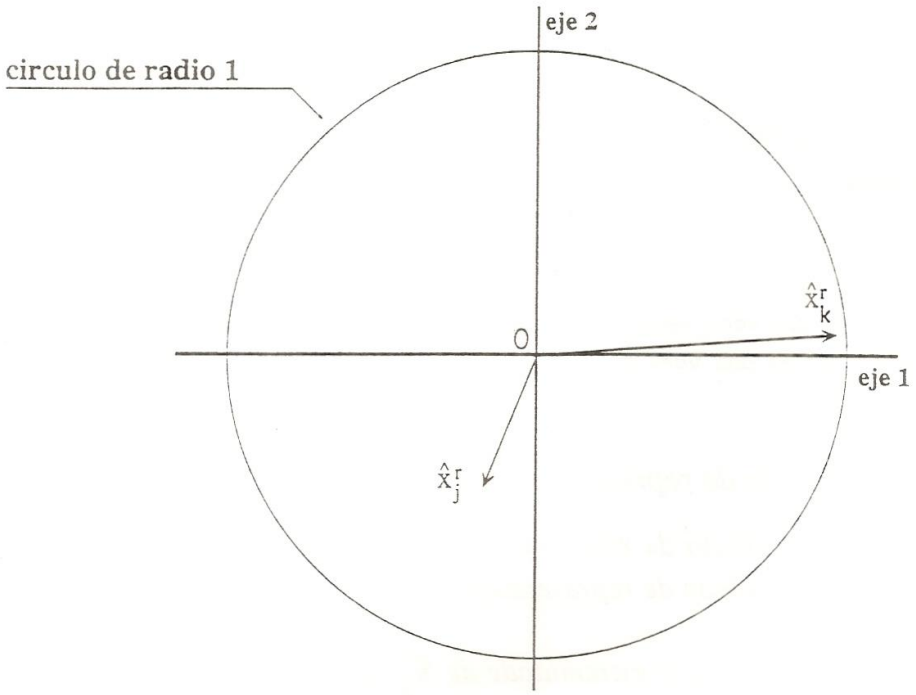


Figura 9

XIII.4. As variáveis no espaço de representação dos indivíduos

- Os indivíduos são representados por pontos. A proximidade entre dois pontos - indivíduos significa a semelhança dos valores das variáveis para esses indivíduos.

- As variáveis são representadas por vetores. A qualidade de representação de uma variável é dada pela proximidade do extremo da mesma ao círculo de raio 1. A correlação entre duas variáveis é representada pelo ângulo que formam os vetores correspondentes.

No espaço de representação dos indivíduos se pode representar os eixos definidos pelos vetores da base ortonormada do referencial de representação das variáveis.

As variáveis iniciais $(x_1,...,x_k)$ *estão representadas pelos eixos definidos pelos vetores da base canônica* $(s_1,..., s_k)$; as componentes principais $(c_1,...,c_k)$ *estão representadas pelos eixos definidos pelos auto-vetores* $(u_1,...,u_k)$.

Os eixos definidos por $(s_1,..., s_k)$ podem ser representados no espaço dos indivíduos.

Se $p=1,\ldots,k$

$$u_k = \sum_{j=1}^{p} s_j \cdot u_{jk} \quad \text{mas também} \quad s_j = \sum_{j=1}^{p} u_k \cdot u_{jk}$$

O vetor projetado no espaço de ordem α, definido por $(u_1,\ldots,u_k)$ *é:*

$$\hat{s}_j = \sum_{j=1}^{\alpha} u_k \cdot u_{jk}.$$

Se esse vetor está bem representado nesse espaço, o eixo definido por ele mesmo pode ser considerado como uma boa representação da j-ésima variável no espaço de representação dos indivíduos.

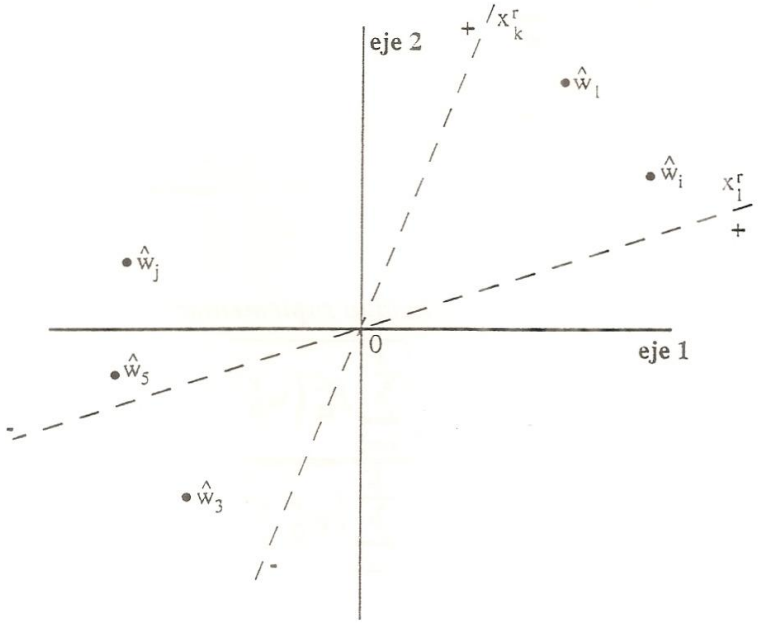


Figura 10

XIII.5. *Indivíduos suplementares ou ilustrativos*

Seja o indivíduo: $w_0 = \{w_{01}, \ldots, w_{0k}\}$

Sendo: $g_j = \{\bar{x}_1, \ldots, \bar{x}_k\}$ e $s_j = \left\{\sqrt{s_{x_1}^2}, \ldots, \sqrt{s_{x_k}^2}\right\}$ *podemos definir o "indivíduo suplementar, centrado e reduzido"*

de termo geral: $w_{0_j}^r = \dfrac{w_{0_j} - g_j}{s_j}$

e $$F_\alpha\left(w_0^r\right) = \sum_{j=1}^{p} u_{\alpha j} w_{0_j}^r$$

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}	V_{15}	V_{16}
URU	72.2	7.8	5916	0.881	2620	3.3	101	21	96	14	5101	163	2.4	47	41	67
CHI	71.8	7.5	5099	0.864	1950	2.9	102	17	93	14	12301	68	2.7	74	26	52
ARG	71	8.7	4295	0.832	2380	1.5	131	30	95	14	3701	115	2.8	62	34	53
VEN	70	6.3	6169	0.824	2560	4.1	99	34	90	9	7001	92	3.2	71	21	56
COL	68.8	7.1	4237	0.77	1260	2.9	106	28	86	30	12301	81	2.7	45	39	68
BRA	65.6	3.9	4718	0.73	2680	3.9	114	59	80	25	10801	96	2.9	25	21	47
EQU	66	5.6	3074	0.646	960	2.7	105	39	84	44	8201	36	3.8	121	33	46
PAR	67.1	4.9	2790	0.641	1090		116	48	88	52	14601	27	4.4	41	11	31
PER	63	6.4	2622	0.592	1100	3.5	87	80	79	30	10401	31	3.7	59	11	53
BOL	54.5	4	1572	0.398	630	2.4	84	89	71	49	15301	29	4.7	101	40	34

Variables:

V_1 : Esperanza de vida al nacimiento, 1990
V_2 : Tiempo medio de escolarización (años), 1990
V_3 : Producto Interior Bruto, per capita (US$), 1990
V_4 : Indice de Desarrollo Humano
V_5 : Producto Nacional Bruto per capita, per capita (US$), 1990
V_6 : Gasto Público en la instrucción (en % del PNB), periodo 1985-90
V_7 : Aporte Calórico Cotidiano (en % de las necesidades normalizadas), 1985-90
V_8 : Taza de mortalidad de menores de 15 años (por 1000 nacidos vivos), 1991
V_9 : Taza de alfabetización de las mujeres (en % de la pop. femenina de edad sup. a 15 años), 1990
V_{10} : Población Rural (en % de la Pop.Total), 1991
V_{11} : Cantidad de habitantes por médico, periodo 1984-89
V_{12} : Cantidad de teléfonos por 1000 hab., periodo 1986-88
V_{13} : Taza de fertilidad, 1991
V_{14} : Monto de la Deuda Total (en % del PNB), 1990
V_{15} : Importancia de la carga de la deuda (en % de las exportaciones de bienes y servicios), 1990
V_{16} : Empleados del Sector Servicios (en % de la Pop. Activa), 1989-91

- Qualidade de representação de um indivíduo suplementar:

$$\frac{\left\|\hat{w}_{0\alpha}\right\|}{\left\|w_0\right\|} = \frac{\sqrt{\sum_{\alpha=1}^{s} F_\alpha^2\left(w_0^r\right)}}{\sqrt{\sum_{j=1}^{K}\left(w_{0_j}^r\right)^2}}$$

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}	V_{15}	V_{16}
URU	72.2	7.8	5916	0.881	2620	3.3	101	21	96	14	510	163	2.4	47	41	67
CHI	71.2	7.5	5099	0.864	1950	2.9	102	17	93	14	1230	68	2.7	74	26	52
ARG	71	8.7	4295	0.832	2380	1.5	131	30	95	14	370	115	2.8	62	34	53
VEN	70	6.3	6169	0.824	2560	4.1	99	34	90	9	700	92	3.2	71	21	56
COL	68.8	7.1	4237	0.77	1260	2.9	106	38	86	30	1230	81	2.7	45	39	68
BRA	65.6	3.9	4718	0.73	2680	3.9	114	59	80	25	1080	96	2.9	25	21	47
EQU	66	5.6	3074	0.646	960	2.7	105	59	84	44	820	36	3.8	121	33	46
PAR	67.1	4.9	2790	0.641	1090	1	116	48	88	52	1460	27	4.4	41	11	31
PER	63	6.4	2622	0.592	1100	3.5	87	80	79	30	1040	31	3.7	59	11	53
BOL	54.5	4	1572	0.398	630	2.4	84	89	71	49	1530	29	4.7	101	40	34

Variables:

V_1 : Esperanza de vida al nacimiento, 1990

V_2 : Tiempo medio de escolarizacion (años), 1990

V_3 : Producto Interior Bruto, per capita (US$), 1990

V_4 : Indice de Desarrollo Humano

V_5 : Producto Nacional Bruto per capita, per capita (US$), 1990

V_6 : Gasto Público en la instrucción (en % del PNB), período 1988-90

V_7 : Aporte Calórico Cotidiano (en % de las necesidades normalizadas), 1988-90

V_8 : Taza de mortalidad de menores de 15 años (por 1000 nacidos vivos), 1991

V_9 : Taza de alfabetización de las mujeres (en % de la pop. femenina de edad sup. a 15 años), 1990

V_{10} : Población Rural (en % de la Pop.Total), 1991

V_{11} : Cantidad de habitantes por médico, período 1984-89

V_{12} : Cantidad de teléfonos por 1000 hab., período 1986-88

V_{13} : Taza de fertilidad, 1991

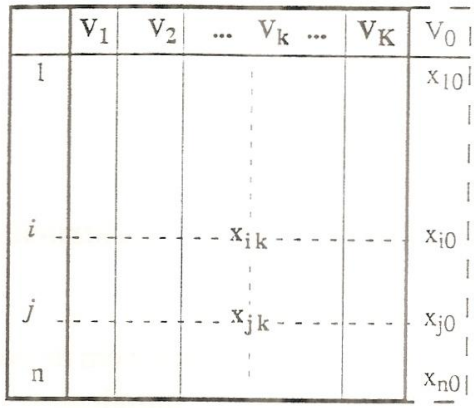
V_{14} : Monto de la Deuda Total (en % del PNB), 1990

V_{15} : Importancia de la carga de la deuda (en % de las exportaciones de bienes y servicios), 1990

V_{16} : Empleados del Sector Servicios (en % de la Pop. Activa), 1989-91

XIII.6. Variáveis suplementares ou ilustrativas

Seja : $x_0 = \begin{Bmatrix} x_{10} \\ x_{20} \\ \vdots \\ \vdots \\ x_{n0} \end{Bmatrix}$



Com : $\bar{x}_0 = \frac{1}{n}\sum_{i=1}^{n} x_{i0}$ y $s_{x_0}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{i0} - \bar{x}_0\right)^2$

pode-se definir: $x_{i0}^r = \frac{x_{i0} - \bar{x}_0}{s_{x_0}}$

e seja $r_{\left(x_0^r, G_\alpha(k)\right)}$ *a correlação da variável suplementar centrada-reduzida com a componente principal* α.

As coordenadas da variável suplementar sobre a base ortonormada $(v_1, \ldots, v_\alpha)$ *são então:*

$$G_\alpha\left(x_0^r\right) = r_{\left(x_0^r, G_\alpha(k)\right)} \quad \forall \alpha = 1, \ldots, K$$

- *Qualidade de representação da variável suplementar:*

$$\left\|\hat{x}_0^r\right\| = \sqrt{\sum_{\alpha=1}^{s}\left(r_{\left(x_0^r, G_\alpha(k)\right)}\right)^2}$$

XIII.7. Fórmula de reconstrução dos dados

A projeção de uma nuvem de pontos sobre os eixos de inércia corresponde a uma mudança de base, introduzindo uma base ortonormada.

A componente $\mathbf{x}_{ik}$ *do i-ésimo indivíduo, na base ortonormada de eixos* $\mathbf{u}_s$, *é dada por:*

$$x_{ik} = \sum_{\alpha=1}^{s} F_\alpha(i)\, u_\alpha(k) = \sum_{\alpha=1}^{s} \frac{F_\alpha(i)\, G_\alpha(k)}{\sqrt{\lambda_\alpha}}$$

Limitando essa soma aos primeiros termos (auo-valores maiores), se obtém valores aproximados de $\mathbf{x}_{ik}$ $\forall$ i y $\forall$ k.

Significado da fórmula de reconstrução dos dados....

$$X = \sum_{\alpha=1}^{s} \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i)\, G'_\alpha(k) = \sum_{\alpha=1}^{s} \sqrt{\lambda_\alpha} v_\alpha\, u'_\alpha$$

A matriz X *é decomposta pela ACP em uma soma de matrizes de faixa de atuação 1.*